

Letter to the Editor

Reversible-Jump Markov Chain Monte Carlo for Quantitative Trait Loci Mapping

Remy van de Ven¹

New South Wales Agriculture, Orange, New South Wales 2800, Australia

Manuscript received December 8, 2003

Accepted for publication March 11, 2004

OVER the past decade there has been a significant increase in the application of Markov chain Monte Carlo (MCMC) methods to modeling data. This can largely be attributed to the renewed interest in Bayesian methods and the increasing power of modern computers. A useful introduction to MCMC methods and their applications is given in GILKS *et al.* (1996). Another area of statistics that has expanded over the past decade is the development of statistical methods to map quantitative trait loci (QTL). It was thus inevitable that MCMC methods would eventually be utilized in QTL mapping. An early example of this approach is by SATAGOPAN *et al.* (1996), who use MCMC methods to map a given number of QTL.

A valuable contribution to MCMC methods came in an article by GREEN (1995) wherein the MCMC methods were extended, using the Metropolis-Hastings algorithm, to include varying dimension problems. This extension, termed reversible-jump MCMC (RJ-MCMC) was soon taken up by the developers of methods to map QTL as it allowed the number of QTL to be included in the model as an unknown. Examples include STEPHENS and FISCH (1998), SILLANPÄÄ and ARJAS (1998), YI and XU (2002), and YI *et al.* (2003).

A problem with application of RJ-MCMC is that care must be taken in determining the acceptance probability for dimension change. As is shown below, some recent publications have determined this value incorrectly. This is partly understandable given that the GREEN (1995) article is rather mathematical. It uses measure theory in its presentation, thus making it less accessible to those less mathematically inclined. To overcome this WAAGEPETERSEN and SORESENSEN (2001) presented an excellent tutorial that largely avoids measure theory. The purpose of this note is to draw attention to an error that appears to be propagating in the literature for the acceptance probability of a dimension change

in the QTL mapping problem. We illustrate the basis of this error with a simplified example.

As our example we assume that we have a data set of n observations such that, for a given s , $\lambda = (\lambda_1, \dots, \lambda_s)$ and covariate values (x_1, x_2, \dots, x_n)

$$Y_i = \sum_{j=1}^s I(x_i > \lambda_j) + \epsilon_i,$$

where $I(\cdot)$ denotes the indicator function and $\epsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$. We assume for simplicity that σ^2 is known to equal 1. Hence the mean is a step function, with s steps at positions λ_j ($j = 1, \dots, s$). The steps can be considered equivalent to QTL locations. For priors on s and λ we let $\Pr(s = k) = q_k$ ($k = 0, \dots, 9$; $\sum_{k=0}^9 q_k = 1$) and let $\lambda|s$ be a simple random sample (without replacement) from $\{0.1 \ 0.1 \ 0.9\}$.

We now use an MCMC algorithm to generate samples from $\pi(\lambda, s|Y, x)$, using a Metropolis-Hastings approach to sample new elements of λ and a RJ-MCMC algorithm to sample s .

Sampling λ_j ($j = 1, \dots, s$): We let the proposal λ_j^* be one of the admissible values for λ_j with equal probability. If ϵ_i and ϵ_i^* are the residuals under the current and proposed (*i.e.*, λ_j^* replacing λ_j) models, respectively, the acceptance probability for this proposal is $\min\{1, \alpha\}$, where

$$\alpha = \frac{\exp\{-0.5 \sum_{i=1}^n (\epsilon_i^*)^2\}}{\exp\{-0.5 \sum_{i=1}^n (\epsilon_i)^2\}}.$$

This step causes no concern.

Sampling s : Let the proposal s^* equal $s + 1$ with probability $p_b(s)$ and equal $s - 1$ with probability $p_d(s) = 1 - p_b(s)$. We let $p_b(0) = 1$, $p_b(k) = 0.5$ ($k = 1, \dots, 8$), and $p_b(9) = 0$.

When deleting a step we choose at random with equal probability one of the steps for removal, whereas when adding a new step, we sample the location of the new step λ^* at random from the admissible steps. One other crucial component of the birth process, overlooked by some, needs to be defined. This is to determine where in the current vector λ the new step λ^* is to be positioned. For our example we choose this position with

¹Address for correspondence: Orange Agricultural Institute, NSW Agriculture, Orange Agricultural Institute Forest Rd., Orange NSW 2800, Australia. E-mail: remy.van.de.ven@agric.nsw.gov.au

equal probability [*i.e.*, $1/(s+1)$ when we have s steps in the current model].

Using this approach for removing or adding a step given s steps currently in the model, the acceptance probability for the proposal is $\min\{1, \alpha\}$, where α is

$$\begin{aligned} \text{Add a step: } \alpha &= \frac{\exp\{-0.5\sum_{i=1}^n(\varepsilon_i^*)^2\}}{\exp\{-0.5\sum_{i=1}^n(\varepsilon_i)^2\}} \times \frac{q_{s+1}}{q_s} \times \frac{p_d(s+1)}{p_b(s)} \\ \text{Drop a step: } \alpha &= \frac{\exp\{-0.5\sum_{i=1}^n(\varepsilon_i^*)^2\}}{\exp\{-0.5\sum_{i=1}^n(\varepsilon_i)^2\}} \times \frac{q_{s-1}}{q_s} \times \frac{p_b(s-1)}{p_d(s)}. \end{aligned}$$

Again ε_i and ε_i^* denote the residuals under the current and proposed models, respectively. The derivations of the acceptance probabilities follow from the results in GREEN (1995) and WAAGEPETERSEN and SORESENSEN (2001).

This expression differs from that of SILLANPÄÄ and ARJAS (1998), YI and XU (2002), and YI *et al.* (2003) who have effectively (ignoring all the other parameters in the QTL mapping context) used the above but with $p_d(s')$ replaced by $p_d(s') \times (1/s')$. That is, they have retained the probability for selecting the particular step for removal but have not included the probability for selecting the position when adding a step. This accounting for the positioning of the added step is essential for balance and reversibility, properties that form the basis of the formulation of the RJ-MCMC algorithm. Two points should be noted here. First, the position of a new step need not be selected with equal probability, but if a position is selected with probability zero, then its counterpart must also to be selected with probability zero. For example, if the location λ^* for a proposed addition were always placed at the end of the current vector λ , then a legitimate RJ-MCMC algorithm can be formulated provided that, when removal of a step is contemplated, the last element of λ is selected with probability one. This situation is covered in WAAGEPETERSEN and SORESENSEN (2001). It has the disadvantage that the resulting chain will not mix as well. It could, however, as suggested in WAAGEPETERSEN and SORESENSEN (2001), be used in conjunction with an additional step being a “shuffle” of the order of the elements of λ . This shuffle would always be accepted and would give rise to an MCMC algorithm as outlined above. The second point is that an MCMC algorithm incorporating the error identified can be shown to be a correct algorithm for a modified model. This modified model has a prior for the distribution of s , the number of steps, proportional to $q_k/k!$, thus penalizing severely models with more steps.

As an aside, another way of thinking about the problem is to consider it as a labeling issue. We need to assign labels to the steps so they are identifiable for selection at the removal stage. So when a step is added or deleted the steps then need to be relabeled. Above we have constrained the relabeling so that those steps not affected directly when sampling s , *i.e.*, not the step

selected for inclusion or removal, retain their same relative order. But there is no need to constrain the labeling to an ordered set of numeric values. We could equally have used any form of label (*e.g.*, personal names, Greek letters) and with a change in dimension reallocate labels, possibly a completely different set for each value of s . Under this approach, if there are $s+1$ steps in the current model, $s+1$ steps can be selected for removal and then there are $s!$ ways to relabel the retained steps. On the other hand, if there are s steps in the current model and a new step is proposed, there are $(s+1)!$ ways to relabel the steps. Hence, using this approach when sampling s again gives, as expected, the same acceptance probability for the proposal.

To illustrate the effect of including and excluding the probability associated with the choice of position we have conducted a limited simulation study. Here we have set $s = 5$ with steps located at 0.1, 0.3, 0.5, 0.6, and 0.8. Also, we let $q_k = 0.1$, $k = 0, \dots, 9$. Ten data sets were generated with each data set composed of 50 observations. The independent vector x was composed of five replicates at each of the points {0.05 (0.1) 0.95}. For each data set an MCMC sample was generated of length 11,000 but with the first 1000 states discarded. In each case the starting value for the chain had $s = 0$ steps in the model. Using the correct acceptance probabilities as given above, the most frequently occurring model (MFOM) was the correct model for five of the simulations, the third MFOM for two simulations, and the fourth MFOM for one. Also, for all but one simulation the MFOM had the minimum residual sum of squares (SS). For the exception, the second MFOM had the minimum residual SS. On the other hand, if for the same data sets samples are generated using the above MCMC algorithm but now excluding the probability associated with the choice of position in the acceptance probability when sampling s , we obtain vastly different outcomes. In all 10 simulations the MFOM contained only one step.

We see therefore that failure to account for the random positioning of a step in the current vector of steps can have a large effect on model selection. The extent of the effect of this error remains to be seen in the QTL mapping context. Indications are that it may not be so influential given that published simulations, based on samples generated using an MCMC algorithm with the error, appear to give reasonable results. Or maybe the error is in the description rather than in the actual computer code.

The author thanks the two referees for constructive comments and Robin Thompson of Rothamsted Research, United Kingdom, for introducing him to this topic.

LITERATURE CITED

- GILKS, W. R., S. RICHARDSON and D. J. SPIEGELHALTER, 1996 Introducing Markov Chain Monte Carlo, pp. 1–19 in *Markov Chain*

- Monte Carlo in Practice*, edited by W. R. GILKS, S. RICHARDSON and D. J. SPIEGELHALTER. Chapman & Hall, London.
- GREEN, P. J., 1995 Reversible jump Markov chain Monte carlo computation and Bayesian model determination. *Biometrika* **82**: 711–713.
- SATAGOPAN, J. M., B. S. YANDELL, M. A. NEWTON and T. C. OSBORN, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**: 805–816.
- SILLANPÄÄ, M. J., and E. ARJAS, 1998 Bayesian mapping of quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.
- STEPHENS, D. A., and R. D. FISCH, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**: 1334–1347.
- WAAGEPETERSEN, R., and D. SORESENSEN, 2001 A tutorial on reversible jump with a view toward applications in QTL-mapping. *Int. Stat. Rev.* **69**: 49–61.
- YI, N., and S. XU, 2002 Mapping quantitative trait loci with epistatic effects. *Genet. Res.* **79**: 185–198.
- YI, N., S. XU and D. B. ALLISON, 2003 Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* **165**: 867–883.

Communicating editor: M. W. FELDMAN

